DOCUMENT RESUME

ED 041 443 88 EM 007 915

AUTHOR Mehlinger, Howard D.; Patrick, John J.
TITLE The Use of "Formative" and "Summative" Evaluation in an Experimental Curriculum Project: A Case in the Practice of Instructional Materials Evaluation.
INSTITUTION Indiana Univ., Bloomington. High School Curriculum Center in Government.
SPONS AGENCY Department of Health, Education, and Welfare, Washington, D.C. National Center for Educational Research and Development.
PUB DATE 6 Mar 70
NOTE 16p.; Paper presented to Annual Meeting of the American Educational Research Association, Minneapolis, Minnesota, March 6, 1970

EDRS PRICE EDRS Price MF-$0.25 HC-$0.90
DESCRIPTORS *Curriculum Development, Curriculum Evaluation, *Evaluation Techniques, *Instructional Materials, Program Evaluation

ABSTRACT

The efforts of a pair of curriculum developers to conduct "formative" and "summative" evaluation of an experimental civics course are described here, along with some of the consequences of their efforts, and a few of the pitfalls they encountered. Formative evaluation refers to those practices that produce data enabling developers to improve their products during the development stage. Summative evaluation refers to an over-all final evaluation of the product with the purpose of producing information useful to the ultimate consumers. Formative evaluation procedures described include pre- and posttesting of student political attitudes, objective testing of performance, open-ended teacher questionnaires, criticism of the course by outside readers, teacher de-briefing sessions, teaching one class by course developers, and site visits to pilot classes. Three instruments were constructed to provide a summative evaluation of the course: a political knowledge test, a political science skills test, and an attitude test. The report describes the plans for administering these tests as well as plans for evaluating the trained versus the untrained teachers and checking teacher and student response to the proposed instructional materials. (JY)

ED041443

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

The Use of "Formative" and "Summative" Evaluation in an Experimental Curriculum Project: A Case in the Practice of Instructional Materials Evaluation

Paper Presented to Annual Meeting of the American Educational Research Association, March 6, 1970

Howard D. Mehlinger and John J. Patrick

Product evaluation presents a number of serious problems to curriculum developers, some that are not resolved by typical evaluation techniques. Scriven's argument that developers consider "formative" and "summative" evaluation stages helps to clarify these problems and offers suggestions to deal with them.[1] This paper describes the efforts by one pair of developers to conduct "formative" and "summative" evaluation of an experimental civics course, some of the consequences of their efforts, and a few of the pitfalls they encountered. The paper makes no attempt to contribute directly to a "theory" of curriculum evaluation. Quite the contrary. By describing a real experience, it will become readily apparent again how wide the gap between theory and practice really is.

For the purpose of this paper "formative evaluation" refers to those practices that produce data enabling developers to improve their products during the development stage. "Summative evaluation" refers to an over-all final evaluation of the product with the purpose to produce information deemed useful to ultimate consumers. While these two stages intersect and even overlap at points, it seems useful for analytical purposes to think of course evaluation as passing sequentially through these two stages.

The "product" referred to in this paper is a two-semester high school social science course entitled "American Political Behavior" under development at Indiana University's High School Curriculum Center in Government. The Government Center, funded by the Cooperative Research Branch of the U.S. Office of Education, was established in 1966 and is sponsored by the Department of Political Science and the School of Education at Indiana University. "American Political Behavior," the

EM 007 915

first course to be developed by the Center, underwent initial pilot trials in 40 schools during 1968-69 and is currently being used in a revised form in 49 different schools. It is important for the purpose of this paper to make clear that the level of project funding has been adequate to support a small, professional and clerical staff but not sufficient to employ professional evaluators. Therefore, the evaluation to be described was planned and carried out by the project directors, the authors of this paper, and fully non-accredited amateur evaluators.[2]

The basic questions that guided the evaluation activities described in this paper are:

1. Can the course "American Political Behavior" be used successfully in the environments provided by typical schools?
2. Can the course be taught as effectively by untrained as by trained teachers?
3. Are there any particular types of students for whom the course seems inappropriate?
4. Can students master the course content?
5. Does the course represent valid political science knowledge and method?
6. Does the course affect students' political attitudes, values, and beliefs in socially desirable ways?
7. Do teachers and students like the course?
8. What types of lessons are most likely to succeed and which are most likely to fail?

Formative Evaluation:

As noted above, formative evaluation refers to those practices that produce data enabling developers to improve their products during the development stage. The following practices were undertaken in an effort to modify and to improve the course "American Political Behavior": pre- and post-testing of student political attitudes; objective testing of student mastery of performance objectives; open-ended teacher questionnaires; criticism of the course by a panel of outside readers;

a meeting at the end of the first year with pilot teachers; teaching of one class by course developers; site visits to pilot classes with interviews of pilot teachers, students, and school administrators. In the paragraphs that follow, each technique will be described; reference will be made to questions the technique sought to answer; changes stimulated by the technique will be cited; and difficulties connected with each technique will be indicated.

Tests of mastery learning. The "American Political Behavior" course is constructed to facilitate mastery learning, the attainment of performance objectives by the majority of students in a particular group.[3] A performance objective is a statement that indicates exactly what a student is able to do as a result of instruction.[4]

Performance objectives are provided with each daily lesson plan in the teacher's guide. Teachers know precisely the purposes of the lesson and can teach to accomplish them. An important element of the instructional strategy is to provide numerous application lessons that enable students to apply knowledge and skills acquired in preceding lessons.

At the end of each instructional sequence, on the average every two weeks, the teachers administered a multiple-choice type examination designed to measure the performance objectives of the material most recently taught. Each item was designed to be a valid measure of one of the objectives. Therefore, theoretically, success on the item represented successful mastery of the objective and the material related to it.

The tests of mastery learning were designed to reveal strengths and weaknesses in the instructional materials. For example, if most students responded correctly to a set of test items pertaining to a performance objective, we assumed that the instructional materials constructed in terms of this performance objective were communicating successfully to students. If most students responded incorrectly to

a set of test items pertaining to a performance objective, we assumed that either the pertinent instructional materials or the test items were flawed and in need of revision. In most instances, a pattern of incorrect student response across different student groups indicated inadequacy of the instructional material and prompted the redesign of particular parts of the course.

It was hoped that the gathering of objective test data from all of the students would be the most powerful and efficient technique for formative evaluation. While it was helpful on several occasions, it was not worth the time, money, and energy given to it. The system was theoretically simple and seemed efficient. However, teachers failed to return tests promptly; some tests were lost; teachers frequently did not check to make certain that answers were recorded in correct places; and students failed to code their tests properly. The result was a gigantic snarl. Special assistants were hired to check individual answer sheets, and computer programmers were hired to try to eliminate some errors by program. The result was an enormous headache and great strain on a limited budget. Probably, we could have accomplished as much by simply asking teachers to record class scores on individual test items. This simple information might have provided better data than we ultimately used.

Teacher questionnaire. At the end of each instructional sequence, approximately ten days, pilot teachers were asked to complete a questionnaire we provided them. Each two to three page questionnaire asked teachers specific questions about individual lessons. It also provided an opportunity for each teacher to comment at length about the course.

The questionnaires frequently were the source of useful tips. We found ideas for the way lessons might be restructured. When the questionnaires revealed that most teachers were having a similar difficulty with a particular segment of the course, we concluded that this portion of the instructional materials probably needed revision.

Panel of outside readers. Two types of readers were used: political science scholars who are specialists in political behavior and specialists in social studies education. The former were used to provide validation of political science content and method in the course; the latter checked us on pedagogical strategies, sequencing of lessons, etc.

Outside readers were used at two different stages. Early drafts of units were sent to readers when the developers were treating concepts that presented special problems for them. When the pilot version of the course was completed, the entire course was read by one political scientist and one social studies specialist who wrote extensive critiques of the material.

The assistance of outside readers was simple to arrange, relatively cheap, and produced excellent results. Ideas for presenting the material were acquired, and some material was entirely rewritten on the basis of the outside assessments. For example, a section on the influence of personality on political behavior was judged particularly weak and has been rewritten to bring it into line with current scholarly views.

End-of-year meeting. In June, 1969, we met approximately one-half of the pilot teachers at a three-day meeting in Bloomington. The purpose of the meeting was to de-brief the teachers on the basis of their experience teaching the "American Political Behavior" course during the 1968-69 academic year. All of these teachers had been trained in a seven-week institute during summer, 1968 prior to teaching the course. The purpose of the summer institute had been less to train them to teach the course than to train them to be critics of the course. In short, they had been trained to become partners in formative evaluation.

At the June meeting, discussion ranged over all elements of the course. The sessions were tape-recorded in order that specific sessions might be replayed if necessary. The session proved to be very valuable, not because it turned up new

problems that had not been recognized earlier, but it tended to confirm the conclusions reached by other evaluation techniques. It was particularly useful to have many teachers present to discuss the course, however, because the complaint of a single teacher often turned out to be less serious than originally believed when it was played out among all the teachers present.

Teachers were particularly warm in their praise of case studies, slide-tape lessons, and the few simulation-games we had provided. Enthusiasm by teachers for the lesson plans we had devised strengthened our resolve to keep them.

Developers' class. Probably the most useful and simple formative evaluation practice is for developers to teach students who are using the experimental course. We gained the permission of local school authorities to establish one section of ninth-graders in a local high school who were our responsibility throughout the school year. By teaching the course, we became instantly aware of serious problems we could repair immediately, without awaiting feedback from other teachers. We were able to make judgments about the readability of the material, pacing, sequencing, etc. When students seemed to lose interest in the course, we were the first to know and were under direct pressure to do something about it.

The principal drawback we found in teaching our own class was the drain on energy and time. When we were meeting our students, we were unable to travel to observe pilot teachers. And we had less time to write. Therefore, this type of evaluation is expensive but probably worth the cost.

Site visits. We were able to visit 30 of the 40 pilot teachers during the first year. When one adds the time required to travel, it is apparent that nearly one-third of the 180-day school year was spent in the field visiting the pilot schools. Site visits are demanding. We talked to the principals, the teachers, and the pilot students at least. Frequently, we were asked to meet other administrators and to speak to the social studies faculty.

Despite the high cost in travel money, time lost, and energy expended, site visits are absolutely essential to the developer. The best way to learn how a course is being taught in a typical classroom is to visit one. Rarely was our course taught exactly as we had conceived it; occasionally it turned out much better than we had imagined it could be; often it was far worse. We found the principal was usually an excellent informant regarding how the course was perceived by the community at large. The students often provided data leading to conclusions that deviated from those derived from test data. It was clear, for example, that students frequently had learned more from the course than test scores had indicated. We learned that in our effort to measure "higher levels" in the Bloom taxonomy, the items became so complex that they were missed because the student could not make sense of the test question. Oral questioning of the students tended to increase our confidence in the course and decrease our confidence in some of the objective test items.

However, site visits tended to support over-all impressions of test data. Where the course was being used with students of low scholastic attainment with limited reading ability, the course was failing. Not surprisingly the course had the greatest success among the highly gifted, academically-inclined students. On the other hand, the course was not only a course for academically able youngsters. It was being mastered by typical ninth-grade youngsters who were reading at eighth- or ninth-grade reading level.

Test of political attitudes. American schools offer courses in civics and government not only because they wish to impart political information, but they also hope to influence students to hold "positive" political values. It is unlikely that any civics course would be accepted by the schools that undermined the attainment by students of socially prescribed "fundamental, American political values." While "American Political Behavior," unlike typical civics and government courses, makes no attempt to preach these values, it certainly intends to support them.

As we were anxious primarily to learn of any "negative" impact the course might have on student political attitudes during the formative evaluation stage, we administered a political attitude instrument as a pre- and post-test to all students taking the pilot course. This political attitude instrument consisted of six sets of Likert-scaled items designed to measure political tolerance, sense of political efficacy, political interest, political trust, support of majority rule practices, and support of political pluralism. This political attitude instrument was used to provide a rough indication of whether or not the course might have a "negative" impact on political attitudes of students. As a whole, the student performance on the political attitude instrument indicated a very slight movement in a "positive" direction on each set of items except the political interest set. Here students showed a very slight decline in political interest. However, as a result of this part of the formative evaluation, we felt no need to massively revise the course for the purpose of reinforcing or creating support for basic democratic political ideals.

Summative Evaluation:

The purpose of summative evaluation is to provide educational decision-makers with evidence about the worth of an educational product, in this instance the "American Political Behavior" course. Before deciding to adopt a course of study, school teachers and administrators should know how the new course performs in terms of particular criteria and how the new course compares with similar products. In order to provide evidence about the worth of a course of study, an evaluator at least must: 1) construct instruments to measure changes in students' behavior toward particular instructional objectives; and 2) administer these evaluational instruments to randomly assigned student groups who have and who have not experienced the experimental instructional materials.

Three instruments have been constructed to measure the impact of the "American Political Behavior" course on students. A *political knowledge* test and a *political science skills* test have been developed to measure student performance in terms of knowledge and cognitive objectives of the course. An *attitude* test has been developed to measure the effect of the course on student political attitudes relating to democratic ideals.

An evaluational instrument that measures knowledge and skill outcomes of instruction must satisfy three basic requirements in order to be valid. First, in order to be a valid test of the relationship of student learning and a course of instruction, test items must fit course objectives. This match between test items and objectives of instruction is the major contributor to the validity of an instrument designed to measure instructional materials. Second, experts must agree on the "right" or "best" answer to each item, if the test is to be considered valid. And third, most students who have not experienced the experimental instructional materials must not be able to respond correctly to the test items.[5]

We need to measure changes in student behavior in order to measure what students have learned as a result of experiencing a particular type of instruction. Tests designed to discriminate individual differences in performance among students do not produce evidence from which one can infer rigorously the relationship of a particular type of instruction to learning. Thus, the standard type of item analysis used in test development does not apply to the development of tests to measure mastery learning. For example, the standard type of item analysis, for the purpose of building tests which measure individual differences, requires the elimination of test items which most students answer correctly or incorrectly. Such items do not discriminate among individual learners. In contrast, the development of tests to measure mastery learning requires the elimination of items which most students answer correctly prior to a particular type of instruction and the retention of

items which most students answer incorrectly prior to instruction. The aim is to build a test which can measure changes in student performance related to particular instruction.[6]

In order to build valid *political knowledge* and *political science skills* tests for use in summative evaluation, we first constructed items that we believe fit our instructional objectives. Next, we sought the aid of political scientists to judge the items, to certify content validity. Then, we administered the tests to students who had not experienced our course in order to determine which items to retain for use in summative evaluation. Items which more than one-half of the "pilot test" students answered correctly were dropped from the instrument, as it was presumed that these items could not help us to measure changes in student performance that were related to experiencing the "American Political Behavior" course.

In order to validly use our tests of *knowledge* and *skills* comparatively, to measure relative performance of groups who have and who have not experienced the "American Political Behavior" course, we wrote items that do not contain jargon peculiar to our course. Students who have not experienced the experimental course should not find it more difficult than students who have experienced the course to read our test items prepared for the summative evaluation. As the tests are free of special terminology, they are more likely to yield differences in understanding and knowledge between different groups of students.

The *attitude* test consists of eight sets of Likert-type items. These eight sets of items, or scales, are designed to measure the following attitudes: political tolerance, sense of political efficacy, political interest, support for majority rule practices, support for political pluralism, political trust, support for practices that equalize opportunities among different socio-economic groups, and support for major institutions of the national government. Collectively these eight sets of items have been devised to yield a rough measure of student support

for "democratic" political practices and basic political institutions of our nation.

Construct validity for the attitude scale has been established through analysis of inter-item correlations. Through this device, the internal consistency of each set of items was established. Items that did not appear to fit, in terms of student responses, with others in a set were dropped from the attitude test.[7]

This spring (1970) we plan to administer the tests of *political knowledge*, *political science skills*, and *political attitudes* to "experimental" and "control" groups in fifteen school systems in five geographical regions. In each case we can claim random assignment of students to an "experimental" group who are taking the "American Political Behavior" course and to a "control" group who are taking another social studies course. The modal grade level of students involved in the summative evaluation is ninth grade, but eighth, tenth, and twelfth grade groups are also present. The student groups represent different socio-economic and ethnic groups. For example, twelfth-graders in the predominately black, inner-city community of Atlanta, Georgia are included in this field trial as are twelfth-grade, middle-class, white students from Eugene, Oregon. Small-town, white, ninth-graders from Mount Vernon, Indiana and white, ninth-graders from the Kansas City metropolitan area are participating in this evaluation of the "product." These examples provide a picture of the range of types of student groups involved in this summative evaluation.

The random assignment of students enables us to claim that the characteristics of the "experimental" and "control" groups are comparable. Thus, we can employ a "post-test only" research design.[8]

In four evaluation sites, the classroom group is the unit of analysis of test results. In each of these situations we have four or five "experimental" groups to be compared with four or five "control" groups. In situations where multiple "experimental" groups can be established, several evaluation experts argue that for

curriculum evaluation the classroom group, rather than the individual students who make up the group, is the most useful unit of analysis.[9] In eleven evaluation sites, we are forced to use the student as the unit of analysis, as we were unable to establish more than one "experimental" and one "control" group.

Through analysis of variance of scores on the *political knowledge* and *political science skills* tests we hope to be able to claim that students who have experienced our course have achieved its basic learning outcomes as specified in performance objectives, and that students in "control" groups have not achieved these outcomes. As other social studies courses, including other civics courses, do not share the knowledge and skill objectives of the "American Political Behavior" course, we are not directly comparing our product with a competing product. In fact, there is no directly competing product, as other civics and government courses represent a legalistic-normative approach to the study of government rather than a social science approach to the study of political activity.

We hope to be able to present evidence to educational decision-makers that our course does communicate effectively to students and that relatively permanent changes in student capabilities have occurred as a result of experiencing the "American Political Behavior" course. Educational decision-makers who value these kinds of changes -- who value the objectives of the "American Political Behavior" course -- are then in a position to decide to utilize the new program. However, educational decision-makers who do not value the kinds of learning outcomes that the new course may effect should not employ the course, even if our summative evaluation indicates that it is an effective product.

Through chi-square and correlational analysis of scores on each of the eight *attitude* scales, we hope to be able to claim, at least, that students who have experienced the "American Political Behavior" course are no more likely to express "negative," or "anti-democratic," political attitudes than are students who have

not experienced the experimental civics course. We would be delighted to be able to claim that our course is related to increased student expression of "democratic" and "positive" political attitudes. And we believe that the new course is likely to reinforce "positive" political attitudes that students bring to the course. However, the performance objectives of our course pertain primarily to cognitive outcomes, to knowledge and skill learnings, not to political attitude outcomes. Thus, we do not anticipate a massive reorientation of student political attitudes to result from experiencing the "American Political Behavior" course.

An additional aspect of the summative evaluation research design calls for the comparison of the instruction of "trained" and "untrained" teachers using the "American Political Behavior" course. "Trained" teachers are those who have experienced, prior to teaching the new course, a special seven-week summer institute taught by the course developers. "Untrained" teachers are those who have not experienced special, intensive instruction prior to teaching the new civics course. The "untrained" teachers have received only written instructions about how to teach the new civics course in a teacher's guide, and they have been given two "position papers" that describe the instructional materials and provide a rationale for the use of the new program.

Among the fifteen school systems involved in the summative evaluation, fifteen "untrained" teachers and nine "trained" teachers are using the "American Political Behavior" course. In three of these school systems, both "trained" and "untrained" teachers are involved in the summative evaluation. We hope that the students of "untrained" teachers perform as well on our instruments of evaluation, relative to their control groups, as the students of "trained" teachers. If this occurs, we can claim that special, intensive instruction is not necessary to prepare a teacher to use the new civics course.

A final feature of our summative evaluation involves the evaluation of the

instructional materials by students and teachers who have used the revised version of the materials in field trials. Both students and teachers in each of the pilot schools who are using the "American Political Behavior" material in field trials will be asked to respond, at the end of the school year, to questions designed to reveal their beliefs about the interest level, relevance, and over-all utility of the new civics course relative to other social studies courses that they have taken. These students and teachers will be asked also to respond positively or negatively to a check list of basic features of the new civics course. These responses can provide a rough indication of student and teacher affect for the new instructional program.

Several difficulties and/or limitations connected with the use of this summative evaluation research design must be indicated. Some of these limitations are minimized through the conduct of several simultaneous, experimental field trials under different conditions. This serves to diminish several possible alternative explanations for the impact of the new course on students that might loom large if the evaluation were conducted only under one set of conditions or only at one site. And it serves to extend the generalizability of our findings.

We face the possibility that several factors other than the instructional materials could account for any successes that are uncovered during the summative evaluations. Factors such as the pedagogical skill or enthusiasm of the teacher, particular learning conditions, the unusual skill or enthusiasm of the students, or community influences could be as important, or more important, than the instructional materials in accounting for successful student performances. However, if each of several, simultaneously conducted field trials produces favorable results under various conditions, our confidence in making claims about the utility of the "American Political Behavior" course will be increased greatly.

Another difficulty connected with summative evaluation is the establishment of

randomly assigned control groups and experimental groups. We were able to establish this condition in only fifteen of the 49 schools that are involved in the field trial of the "American Political Behavior" course.

Still another difficulty involves the need to obtain accurate information about the prior curricula experience of students involved in the evaluation and about special or unusual conditions affecting the learning environment. For example, students in one of our experimental sites have been involved in field trials of the anthropology and geography project materials prior to experiencing our course. This prior experience is likely to affect their performance in our course. This kind of information about the context of each field trial is necessary in order to interpret satisfactorily the findings of the summative evaluation.

Because of the large number of variables and because of the many difficulties involved in conducting summative evaluation, we cannot be certain that successful student performance results directly and entirely from experiencing the "American Political Behavior" course. But through the use of the summative evaluation procedures described in this paper, we can claim that particular students do, or do not, attain specific learning outcomes that are integrally involved in the new course. And we increase the probability that our claims about the efficacy and/or weaknesses of the course are accurate. These results of summative evaluation provide educational decision-makers with grounds for deciding whether or not to utilize the new instructional materials.

This description of the "formative" and "summative" evaluation of the "product" of a social studies curriculum project reveals some of the pitfalls, limitations, and fruitful possibilities involved in this two-stage evaluation process. Hopefully, this recounting can serve others who are interested in the challenges of instructional materials development.

FOOTNOTES

[1]Scriven, Michael. "The Methodology of Evaluation." In Perspectives of Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally and Company, 1967, pp. 39-83.

[2]The following articles provide descriptions of the course in "American Political Behavior" and the assumptions of the course developers. Mehlinger, Howard. The Study of American Political Behavior. Bloomington: Indiana University, 1967, unpublished paper; Patrick, John J. "Teaching High School Students about American Political Behavior." The North Central Association Quarterly 43:234-242, Fall, 1968.

[3]Bloom, Benjamin S. "Learning for Mastery." UCLA-CSEIP. Evaluation Comment, May, 1968.

[4]Gagne, Robert M. "Curriculum Research and the Promotion of Learning." In Perspectives of Curriculum Evaluation. AERA Monograph Series on Curriculum Evaluation. Chicago: Rand McNally and Company, 1967, p. 21.

[5]Scriven, op. cit., pp. 45-60; Wittrock, M.C. "The Evaluation of Instruction: Cause and Effect Relations in Naturalistic Data." UCLA-CSEIP. Evaluation Comment, May, 1969.

[6]Ibid., pp. 4-5.

[7]Oppenheim, A.N. Questionnaire Design and Attitude Measurement. New York: Basic Books, Inc., 1966, pp. 133-143.

[8]Popham, W. James. Simplified Designs for School Research, Inglewood, California: Southwest Regional Laboratory for Educational Research and Development, October, 1967; Kerlinger, Fred N. Foundations of Behavioral Research. New York: Holt, Rinehart and Winston, Inc., 1966, pp. 301-321.

[9]Popham, op. cit., p. 11; Baker, Robert L. "Curriculum Evaluation." Review of Educational Research 39:341, June, 1969.